1. What is Machine Learning (ML)?

Machine Learning is a branch of **Artificial Intelligence (AI)** that allows computers to **learn from data and improve automatically** without being explicitly programmed.

How it works:

- The system is given **data**
- It finds **patterns**
- It makes **predictions or decisions**
- Performance improves as **more data** is used

Example:

- Email spam filtering
- YouTube video recommendations
- Face recognition

Types of Machine Learning:

1. **Supervised Learning** – uses labeled data (e.g., predicting marks)
2. **Unsupervised Learning** – finds patterns in unlabeled data (e.g., clustering)
3. **Reinforcement Learning** – learns through rewards and penalties (e.g., game AI)

2. Analytics vs Data Science

Aspect	**Analytics**	**Data Science**
Focus	Analyzing past data	Predicting and discovering insights
Scope	Narrow	Broad
Tools	Excel, SQL, BI tools	Python, R, ML, AI
Output	Reports & dashboards	Predictions & models
Goal	"What happened?"	"What will happen & why?"

➡️ **Analytics is a part of Data Science**, but Data Science includes ML, AI, and advanced modeling.

3. Value Chain (Analytics / Data Science Value Chain)

The **value chain** shows how raw data becomes useful business value.

Steps:

1. **Data Collection** – from sensors, websites, apps
2. **Data Storage** – databases, cloud
3. **Data Cleaning** – remove errors, duplicates
4. **Data Analysis** – statistics, analytics
5. **Model Building** – ML/DL models
6. **Insights & Decisions** – business actions
7. **Value Creation** – profit, efficiency, growth

---

4. Types of Analytics

1. Descriptive Analytics

- **What happened?**
- Example: Monthly sales report

2. Diagnostic Analytics

- **Why did it happen?**
- Example: Drop in sales due to price increase

3. Predictive Analytics

- **What will happen?**
- Example: Forecasting future demand

4. Prescriptive Analytics

- **What should we do?**
- Example: Suggest best pricing strategy

5. Lifecycle Probability (Model Probability Lifecycle)

This refers to how **probability-based predictions** evolve during model usage.

Stages:

1. **Data Input** – historical data

2. **Training Phase** – model learns probabilities
3. **Validation –** test accuracy
4. **Prediction** – outputs probability (e.g., 80% chance of rain)
5. **Monitoring** – check performance over time
6. **Re-training** – update with new data

6. Analytics Project Lifecycle

A standard analytics project follows these steps:

1. **Problem Definition**
   - Understand business objective
2. **Data Collection**
   - Gather relevant data
3. **Data Preparation**
   - Clean and preprocess data
4. **Exploratory Data Analysis (EDA)**
   - Find trends and patterns
5. **Model Building**
   - Apply ML algorithms
6. **Evaluation**
   - Measure accuracy, performance
7. **Deployment**
   - Use model in real-world system
8. **Monitoring & Maintenance**
   - Update model regularly

---

7. Advantage of Deep Learning over Machine Learning

Feature	Machine Learning	Deep Learning
Feature Engineering	Manual	Automatic

Feature	Machine Learning	Deep Learning
Data Handling	Small to medium	Very large datasets
Accuracy	Good	Very high
Complexity	Low to medium	Very high
Best for	Structured data	Images, audio, video

---

8. Reasons for Deep Learning

Deep Learning is preferred because:

- Handles **huge amounts of data**
- Works well with **unstructured data**
- Learns **complex patterns**
- Minimal human intervention
- High accuracy in vision & speech
- Powered by **GPUs and cloud computing**

---

9. Real-Life Use Cases of Deep Learning

Healthcare

- Disease detection from X-rays
- Medical image analysis

Transportation

- Self-driving cars
- Traffic prediction

Finance

- Fraud detection
- Stock price prediction

Entertainment

- Netflix recommendations
- Game AI

Security

- Face recognition
- Voice authentication

10. Review of Machine Learning

Strengths:

- Automates decision-making
- Improves accuracy
- Learns from data
- Scales well

Limitations:

- Needs quality data
- Can be biased
- Requires computational power
- Hard to explain complex models

Future Scope:

- Integration with AI
- Smarter automation
- Human-AI collaboration

1. Basis of Data Categorization

Data is categorized based on **how it is measured, structured, and used**.

Common bases of categorization:

1. **Nature of data** – qualitative or quantitative
2. **Structure** – structured, semi-structured, unstructured
3. **Source –** primary or secondary
4. **Time** – historical or real-time
5. **Usage –** transactional or analytical

---

2. Types of Data

1. Qualitative Data (Categorical)

- Descriptive, non-numerical
- Example: Gender, color, feedback

Sub-types:

- **Nominal** – no order (e.g., blood group)
- **Ordinal** – ordered (e.g., ratings)

2. Quantitative Data (Numerical)

- Numeric and measurable

Sub-types:

- **Discrete** – countable (e.g., number of students)
- **Continuous** – measurable (e.g., height, weight)

---

3. Data Collection Types

1. Primary Data Collection

Data collected **first-hand** for a specific purpose.

Methods:

- Surveys
- Interviews

- Observations
- Experiments

2. Secondary Data Collection

Data collected **by others** and reused.

Sources:

- Government records
- Company databases
- Research papers
- Websites

---

4. Forms of Data & Sources

Forms of Data:

1. **Structured Data**
   - Tables, rows, columns
   - Example: Databases
2. **Semi-Structured Data**
   - Has tags or markers
   - Example: XML, JSON
3. **Unstructured Data**
   - No fixed format
   - Example: Images, videos, emails

Data Sources:

- Sensors & IoT devices
- Social media platforms
- Enterprise systems (ERP, CRM)
- Web applications
- Mobile apps

---

5. Data Quality & Changes

Data Quality

Refers to how **accurate, complete, consistent, and reliable** data is.

Data Changes

Data can change due to:

- New data entries
- Updates or corrections
- Business rule changes
- System migrations

Continuous monitoring is required to maintain quality.

---

6. Data Quality Issues

Common problems include:

- **Missing values**
- **Duplicate data**
- **Inconsistent formats**
- **Outdated information**
- **Incorrect data**
- **Data entry errors**

These issues can lead to **wrong analysis and poor decisions**.

---

7. Data Quality Story

A **Data Quality Story** explains how poor data affects outcomes.

Example:

A hospital system stores patient age incorrectly.

- Wrong data → Wrong diagnosis
- Wrong diagnosis → Wrong treatment
- Result → Risk to patient life

➡️ **Conclusion:** Good decisions require good data.

---

8. What is Data Architecture?

Data Architecture is the **blueprint** that defines:

- How data is collected
- How it is stored
- How it flows
- How it is used across an organization

It ensures **data consistency, security, and accessibility**.

---

9. Components of Data Architecture

1. Data Sources

- Databases, sensors, applications

2. Data Storage

- Data warehouses
- Data lakes
- Cloud storage

3. Data Processing

- ETL (Extract, Transform, Load)
- Data pipelines

4. Data Integration

- Combining data from multiple sources

5. Data Governance

- Policies, security, compliance

6. Data Consumption

- Reports, dashboards, analytics

---

10. OLTP vs OLAP

Feature	OLTP	OLAP
Full Form	Online Transaction Processing	Online Analytical Processing
Purpose	Day-to-day operations	Analysis & reporting
Data	Current, detailed	Historical, summarized
Queries	Simple & fast	Complex
Example	Banking transactions	Sales trend analysis

---

11. How is Data Stored?

1. Databases

- Relational (MySQL, Oracle)
- NoSQL (MongoDB)

2. Data Warehouses

- For analytical data
- Example: Amazon Redshift

3. Data Lakes

- Stores raw data
- Supports all data types

4. Cloud Storage

- Scalable and cost-effective
- Example: AWS S3, Google Cloud

5. File Systems

- CSV, Excel, text files